Article

JoLIS

Journal of Librarianship and Information Science
1–15

DOI: 10.1177/0961000618759413
journals.sagepub.com/home/lis
SAGE

Mining location information from life- and earth-sciences studies to facilitate knowledge discovery

Jason W. Karl
University of Idaho, Moscow, Idaho, USA

Abstract

Location information in published studies represents an untapped resource for literature discovery, applicable to a range of domains. The ability to easily discover scientific articles from specific places, nearby locales, or similar (but geographically separate) areas worldwide is important for advancing science and addressing global sustainability challenges. However, the thematic and not geographic nature of current search tools makes location-based searches challenging and inefficient. Manually geolocating studies is labor intensive, and place-name recognition algorithms have performed poorly due to prevalence of irrelevant place names in scientific articles. These challenges have hindered past efforts to create map-based literature search tools. Thus, automated approaches are needed to sustain article georeferencing efforts. Common pattern-matching algorithms (parsers) can be used to identify and extract geographic coordinates from the text of published articles. Pattern-matching algorithms (geoparsers) were developed using regular expressions and lexical parsing and tested their performance against sets of full-text articles from multiple journals that were manually scanned for coordinates. Both geoparsers performed well at recognizing and extracting coordinates from articles with accuracy ranging from 85.1% to 100%, and the lexical geoparser performing marginally better. Omission errors (i.e. missed coordinates) were 0% to 14.9% for the regular expression geoparser and 0% to 10.3% for the lexical geoparser. Only a single commission error (i.e. erroneous coordinate) was encountered with the lexical geoparser. The ability to automatically identify and extract location information from published studies opens new possibilities for transforming scientific literature discovery and supporting novel research.

Keywords

Automated georeferencing, geographic coordinates, geotagging, literature discovery, parser, publishing standards

Introduction

Location information in published studies represents a huge and largely untapped resource for knowledge discovery that is applicable to a wide range of domains including earth sciences, ecology, conservation, environmental science, and human health (Karl et al., 2013; Maggio et al., 2017). The ability to quickly and efficiently discover existing knowledge from specific places, nearby locales, or similar areas worldwide is important for supporting advances in science as well as for addressing global conservation and sustainability challenges (Karl et al., 2013). Searching for scientific knowledge geographically as well as thematically can improve the accessibility of potentially relevant research and enable knowledge discovery from environmentally similar but geographically separated areas (Karl et al., 2013; Page, 2010). Geographic literature searching can also promote syntheses and meta-analyses (Hughes et al., 2002; Magliocca et al., 2014; Van Vliet et al., 2012), improve understanding of environmental patterns (Jetz et al., 2012), and facilitate evaluations of bias in scientific knowledge (Fisher et al., 2011; Martin et al., 2012).

However, discovering relevant knowledge about specific places has typically involved researchers working from their own knowledge, querying professional networks, and tedious searching (Zimmerman, 2007) because current search tools for scientific literature are largely thematically and not geographically oriented, making location-based searches challenging and inefficient (Karl et al., 2013). Searches for literature from specific places using place names terms often yields many results from outside

Corresponding author:
Jason W Karl, Department of Forest, Rangeland, and Fire Sciences, University of Idaho, 875 Perimeter Drive MS 1135, Moscow, Idaho 83844-1135, USA.
Email: jkarl@uidaho.edu

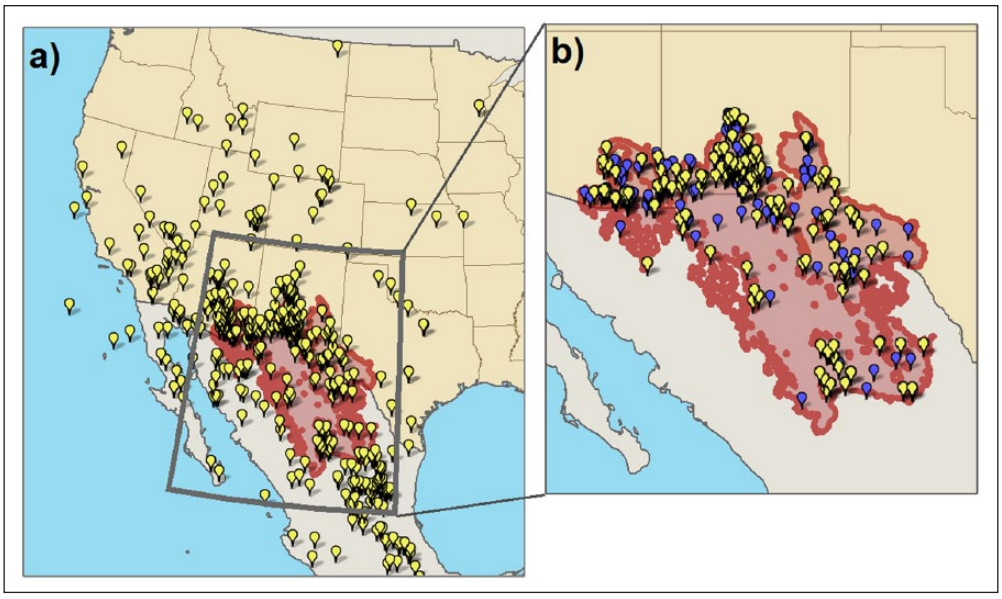

Figure 1. Articles resulting from a Web of Science search for ecology articles using the geographic search term "Chihuahuan Desert".
Note: These were manually georeferenced (yellow pins, Chihuahuan Desert area shown in red) to illustrate the shortcomings of literature discovery using geographic place names. Only 33% of the over 800 articles georeferenced occurred in the Chihuahuan Desert (a). Additionally, a geographic search of the JournalMap database found many articles not returned by the original keyword search (b, blue pins).

the desired area (Figure 1). Without the use of an ontology to link nested geographic concepts (e.g. Jones et al., 2002), relevant sources that did not use the same geographic descriptions may also be missed. Additionally, with only thematic searching it is not possible to discover literature from places that share a similar geographic context. One solution for these problems is to create geographic (i.e. map-based) literature search tools that use databases of georeferenced literature.

Background and statement of need

The value of georeferenced literature databases has been demonstrated for many fields including ecological research (Martin et al., 2012), conservation (Schmitt and Butler, 2012), global change (Schmill et al., 2014), biodiversity (Page, 2011), and infectious disease (Hendrickx et al., 2010). However, in most cases, these have been labor-intensive efforts that required skimming each article to determine the study location and extent. Such a high time and effort investment has relegated large-scale geolocating of historic articles to "georeferenced bibliographies" of specific topics or places (e.g. Schmitt and Butler, 2012) and may contribute to the difficulty in sustaining these efforts in the long term (e.g. Hendrickx et al., 2010). Additionally, until a very large number of articles relative to a topic are georeferenced and searchable by location, the potential to analyze geographic patterns and explore spatial relationships is limited. Thus, the creation, longevity, and ultimate usefulness of georeferenced literature databases could be greatly improved if location information could be reliably and easily extracted from published studies in an automated manner.

Several new search tools have been developed that leverage the locations of scientific field research to provide for map-based searching and geographic similarity, and this has begun to change the thematic-only literature search paradigm. For example, the data repository Pangaea (http://www.pangaea.de) provides mapped locations of datasets that can be tied back to published articles. The USGS Science Base (https://www.sciencebase.gov) database assigns geographic coordinates or bounding boxes to Government Reports and articles published by agency scientists. Article abstracting services like GeoRef (http://www.americangeosciences.org/georef/about-georef-database) and CAB Abstracts (http://www.cabdirect.org/) assign geographic locations (but not always based on coordinates) to articles. The GLOBE project (Schmill et al., 2014; http://globe.umbc.edu/) maps published case studies and provides analytic tools to support meta analyses of land use change around the world. JournalMap (Karl et al., 2013; http://www.journalmap.org) provides a map-based search interface for georeferenced journal articles and an API for embedding article maps. BioStor (http://www.biostor.org) extracts geographic locations from historic articles in the Biodiversity Heritage Library and provides article-level maps of specimen locations (Page, 2011). Of the existing examples, most assign locations to published articles either via their source data (e.g. Pangaea), via self-reporting by the authors or manually (e.g. ScienceBase, GLOBE, CAB Abstracts). Currently though, only JournalMap and BioStor attempt to automatically geolocate already-published

studies from location information reported in the article text. However, these approaches will continue to be of limited utility until the total amount of georeferenced literature is greatly increased, and this will hinge on developing techniques for rapidly and accurately georeferencing existing literature.

Place name descriptions, maps, and geographic coordinates are common ways of describing study area locations in scientific articles. While not all published studies provide coordinates to demark the study's location, almost all studies that pertain to real areas on Earth contain a narrative description of the study area. Shapiro and Baldi (2012) found that reported geographic coordinates did not match the place name description of studies in 16% of articles they reviewed, suggesting that place name geolocation might be more accurate than relying solely on geographic coordinates. Also, it is possible using search algorithms or natural language processing to automatically detect, extract, and assign geographic coordinates to place names contained in unstructured text like a scientific article (Jones and Purves, 2008). Such approaches rely on either a simple gazetteer (Amitay et al., 2004; Smith and Crane, 2001) or ontology (Jones et al., 2002; Volz et al., 2007) of place names and their assigned coordinate values.

While many different text-mining approaches have been proposed for detecting and geocoding place-name locations (e.g. Borges et al., 2011; Leidner, 2007; Lieberman et al., 2010; Lieberman and Samet, 2011) and may have value for other applications, simply applying these techniques to geotagging scientific articles for the purpose of mapping the location(s) where the study originated would yield many irrelevant place names and obscure the actual study location for several reasons. First, scientific articles often contain many irrelevant place names. While some of these erroneous locations can be easily filtered out (e.g. place names in author affiliations or reference sections), in many cases it is difficult to discriminate between an irrelevant place name and one that describes the location where the study occurred. For example, place names commonly occur in ecological studies when referencing other studies (e.g. "Gillan et al. (2013) found that Greater Sage-grouse avoided powerlines up to a distance of 600m in west-central Idaho."). Place names can also frequently occur in species common names (e.g. Idaho fescue, Canada goose). Another source of irrelevant place names in scientific articles is the convention of providing the company address (typically city, state, and/or country) when citing the use of commercial products. Second, many authors report non-standard (e.g. "the Bureau of Land Management's Wildhorse Allotment in southern Idaho") or imprecise (e.g. "a horse farm in Kentucky") place names when describing their study areas. Third, there is no standardization for where and how study locations are described in an article. Thus, there is much ambiguity in textual descriptions of place names in published documents (Lieberman et al., 2010), and given existing tools could be an error-prone technique for creating georeferenced literature databases.

Alternatively, many published articles also describe study area locations using geographic coordinates, and this more precise geographic information could be used to overcome challenges of georeferencing literature from place names. Incidence of geographic coordinate reporting in studies varies by discipline, with the highest rates (approximately half of articles reporting geographic coordinates) seen in place-based sciences like ecology and agriculture (Karl et al., 2013). Pattern-matching algorithms (i.e. parsers) have been successfully used to identify, extract, and standardize geographic coordinates from the text of scientific articles (e.g. Karl et al., 2013; Page, 2010). Thus, the ability to geotag articles quickly and easily from geographic coordinates could greatly increase the usefulness of geographic literature searching by quickly and accurately georeferencing articles and serving as validation to place-name georeferencing approaches. However, identifying, extracting, and standardizing this location information is not trivial because few conventions exist for reporting study locations and article formatting varies widely (Karl et al., 2013). Consequently, performance of common parsing approaches for detecting and translating geographic coordinates needs to be evaluated in the context of building georeferenced literature databases and search tools.

Objective

The objective of this paper was to describe and test regular expression and lexical parsers for automatically geotagging scientific articles based on reported geographic coordinates. Development of the two parsers and their performance against a training set of coordinates are first described, followed by testing the coordinate parsers against independent sets of full-text articles from multiple journals that were manually scanned for coordinates. The potential merits, limitations, and applications of these approaches to location mining are then explored. Finally, standards are discussed for how and when study area locations should be reported in a scientific article, and how to improve the ability to capture geographic information from articles and enable geographic searching and literature visualization.

Materials and methods

Parsing

Parsing is the process of recognizing a string of characters as belonging to a specific class or language by breaking the string down into sets of symbols and analyzing each set against predefined rules (Grune and Jacobs, 1990). The

rules a parser (i.e. the computer algorithm that performs parsing) relies upon are expressed as a grammar that defines the forms of strings that can be recognized based on a collection of basic components (e.g. alphanumeric characters, punctuation) and syntax. A parser does not describe the meaning of a string, but simply identifies and standardizes the formatting of strings that belong to a particular grammar. These standardized strings can then be passed to other algorithms for further processing and interpretation. The theory behind parsers and grammars in computer science is well developed as they form the basis for interpreting computer languages into machine directions (see Aho et al., 1986). Parsers, however, can also be used in a pattern-matching context to identify and extract strings of a specific form like geographic coordinates (Ford, 2004). While there are many different approaches to parsing and styles of parsers, this paper considers two different parsers with unique approaches to defining their grammars: regular expressions and lexical parsing.

Parsing with regular expressions

Regular expressions were first developed in the 1950s as a notation system for languages and were first implemented in computer systems in the 1960s for defining search patterns (Johnson et al., 1968; Thompson, 1968), and see wide use today in many computing applications. Regular expressions consist of a sequence of characters that describe a search pattern (Friedl, 2006). A regular expression processor (a component of almost all computer languages) passes the pattern definition over a string or block of text to be searched and any segments matching the search pattern are identified and can be extracted. For example, running the regular expression "the" over the sentence "We went to see the theater to see a movie" would return two instances of search string – "the" and "theater". Because they are frequently used to extract data from large files, regular expressions can be thought of as an implicit expression of the format of a type of data. Regular expressions can consist of literal characters (e.g. letters, numbers) as well as special "metacharacters" that define parsing concepts like alternation, repetition, grouping, and wildcards (i.e. any character). Extending the example above, adding the "\b" metacharacter for word boundaries to the regular expression (i.e. "\bthe\b") would restrict the pattern to only instances of the three characters occurring together as a single word and return one instance for the actual word "the". While regular expressions do not formally define a parsing grammar, the use of named groups allows different parts of a recognized pattern to be split out and labeled for further processing. For a more thorough treatment of regular expressions, see Friedl (2006).

Regular expressions are a compact and fast way to define simple search patterns. However, owing to their compact nature, regular expressions can be difficult to read for complex patterns – making them difficult to troubleshoot and update.

Lexical parsing

Lexical parsing has long been used in computer science applications, but has not seen widespread use for identifying and extracting geographic locations from unstructured text documents. Hence the need to establish the utility of this technique and assess its performance relative to typical regular expression applications for geoparsing. Lexical parsing begins with converting the characters of a string or document into a sequence of tokens (i.e. words, phrases, or other meaningful elements) that have an ascribed meaning. The parsing algorithm then applies the grammar syntax to the set of tokens to identify which ones belong to the grammar (e.g. are geographic coordinates). The grammar for lexical parsing builds basic "letter" and "word" concepts into more complex structures which are then linked together. For recognizing geographic coordinates, basic elements of digits and symbols (e.g. degree signs) are grouped together into larger components like degrees, minutes, seconds, or hemisphere designations. These components are further assembled into the latitude or longitude parts of the coordinate, and these parts are finally assembled into the coordinate pair. A lexical grammar allows for specification of how the elements are assembled and any validation rules that must be part of a valid coordinate (e.g. minutes or seconds must be a positive number less than 60). While this results in grammars that are more verbose than regular expressions, they are much easier to read and troubleshoot. For the purposes of geoparsing journal articles, a lexical parser would break the entire article body into tokens and then parse them according to the coordinate parser's grammar syntax. This tokenization of large documents contributes to lexical parsers being slower and computationally more intensive than the pattern-matching progression of regular expressions.

Anatomy of a coordinate: Building a geoparsing grammar

To develop a parser for geographic coordinates, the basic structure of a coordinate was first defined and rules determined for which elements of the coordinate were required, how order of the elements could vary, and valid ranges for the elements. It is possible to define a coordinate parser for any type of structured expression of a geographic coordinate (e.g. points, lines, bounding boxes, or polygons in a variety of coordinate systems). Through the JournalMap project (Karl et al., 2013), it was found that over 90% of articles that contain coordinates report coordinates for points using a geographic coordinate system (i.e. degrees latitude and longitude). Less than 10% of studies used coordinates to describe bounding boxes or other feature

Table 1. Coordinate parts and value ranges used in developing the coordinate parsers.

Coordinate Element	Type	State	Value Range*	Examples
Coordinate pair	Character string	Required		16° 48.16' N, 88° 04.94' W
Latitude coordinate part	Character string	Required		16° 48.16' N
Longitude coordinate part	Character string	Required		88° 04.94' W
Coordinate pair separator	Character symbol or string	Optional	Literal set (5)	„; and by
Latitude hemisphere	Character symbol or string	Optional	Literal set (4)	N S North South
Longitude hemisphere	Character or word	Optional	Literal set (4)	E W East West
Negative sign	Character symbol	Optional	Literal set (6)	- –
Degrees	Real number	Required	[-90, 90] for latitude; [-180, 180] for longitude	16
Degree symbol	Character symbol or string	Required	Literal set (8)	° º deg degrees
Minutes	Integer or decimal number	Optional	(0, 60)	48.16
Minutes symbol	Character symbol or word	Optional	Literal set (9)	‘ ' min minutes
Seconds	Integer or decimal number	Optional	(0, 60)	52.23
Seconds symbol	Character symbol or word	Optional	Literal set (18)	“ " sec seconds

*A literal set consists of a specific set of character symbols or strings (i.e. words) which may satisfy the requirement of that coordinate element. Only one of the set is possible in each instance. When matching subsets of a coordinate string to a literal set, case was not considered. The number in parentheses is the total number of options or alternatives acceptable for that literal set.

types, and only 5% of studies reported coordinates using a different coordinate system (e.g. Universal Transverse Mercator). Thus, for the purposes of this study, inquiry was limited to recognizing points in degrees latitude and longitude. Additional coordinate parsers are in development that expand on the concepts described below for bounding boxes and other coordinate systems.

The syntactic structure of a coordinate was described in a parse tree (Figure 2) and as a set of coordinate parts and value ranges (Table 1). For this study, the requirements to be considered as a valid coordinate for the coordinate parsers were that the coordinate pair must: (1) be in a geographic coordinate system (i.e. expressed as some combination or fraction of degrees, minutes, and seconds), (2) consist of at least degrees latitude and degrees longitude, and (3) contain a degree sign for both the latitude and longitude parts of the coordinate pair. The degree sign requirement was necessary because many pairs of numbers in a scientific article (e.g. morphologic measurements for a species) can occur within the same value ranges as coordinates and cause the coordinate parser to return false positives. Semantic cues for different parts of a coordinate are also specified in Table 1.

Coordinate parser development

Each coordinate parser was developed using the same test set of locations from the JournalMap database consisting of a set of articles having geographic coordinates were also manually georeferenced, thus providing a determination of the coordinate's standardized value independent of the coordinate parsers. For the purposes of this paper only geographic coordinates using a combination of decimal degrees, minutes, and seconds (e.g. the following would all be valid coordinates: 41.326° N by 83.112°W; 41°44.48′ N by 83°28.51′ W; or 41°41′49″S by 128°16′48″E) were used. While a coordinate parser could be written for any geographic coordinate system, results from the JournalMap project have shown that alternative coordinate systems like Universal Transverse Mercator (UTM) or polar stereographic account for less than 5% of coordinates reported in ecological literature. For each article the verbatim coordinate strings were copied from the article to use for training and testing the coordinate parsers. Locations with obvious errors (e.g. latitudes greater than 90º) were excluded (approximately 0.8% of all locations). An exception to this was when both a hemisphere designation and a negative sign was included in a coordinate part (e.g. -114.34ºW). In these cases, I assumed that the hemisphere designation provided was correct and ignored the negative sign. Coordinates representing bounding boxes, ranges or extents (e.g. between 46°33.85′ N, 90°25.06′ W and 41°44.48′ N, 83°28.51′ W) were also excluded. This process yielded 3159 unique locations from 795 articles. Both coordinate parsers were written in Python version 2.7.10 (https://www.python.org).

Regular expression coordinate parser

The regular expression coordinate parser was written using Python's built-in *re* library (full code at https://github.com/JournalMap/JLIS_Geoparsers). I relied on the original concept of a general-purpose regular expression for extracting geographic coordinates defined in GeoLucidate 0.3 (K. Raschke, 2010; https://github.com/kurtraschke/geolucidate). This original code was heavily modified to be able to recognize many more types and formats of geographic coordinates. The final regular expression

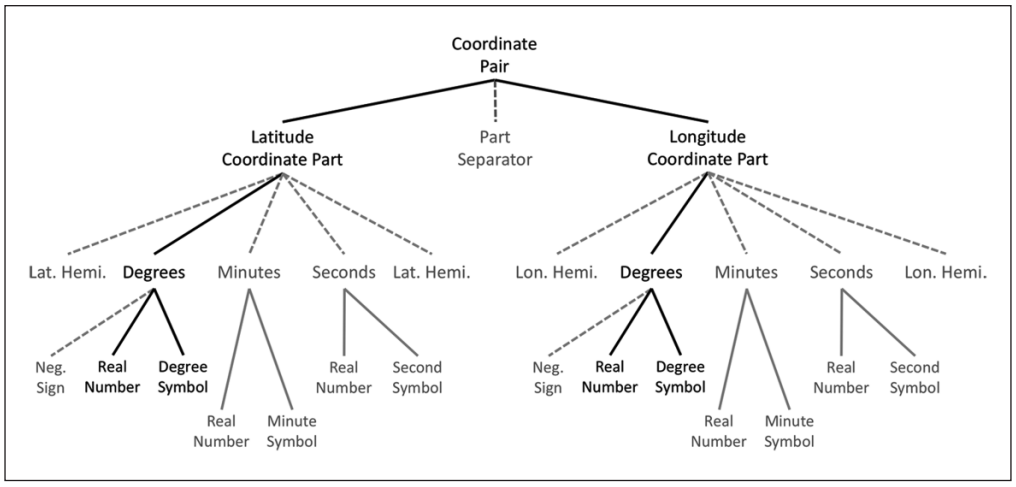

Figure 2. Simplified parse tree for the coordinate parser representing the grammar to define a geographic coordinate.
Note: Elements shown in black are required for recognizing a legitimate coordinate. Elements shown in gray with a dashed gray line are optional. Elements in gray with a solid gray line are conditional (i.e. if minutes are given, they must consist of a real number and a minute symbol). Descriptions of the elements and examples of valid values are given in Table 1.

coordinate parser is described in the syntax diagram in Figure 3. Regular expressions parse strings from left to right, collecting characters that meet the defined syntax definitions. Optional characters allow for components of a coordinate (e.g. symbols for minutes or seconds) or stylistic elements (e.g., spaces) to be present or not or to be in different locations within the coordinate (e.g. hemisphere designations). To be considered a coordinate by the regular expression, the parsed string must contain all of the required elements. Portions of a regular expression can be assigned names which can then be used to extract and transform (i.e. convert to a standard format) specific parts of the coordinate.

Lexical coordinate parser

The lexical coordinate parser was written in PyParsing 2.1.5 (McGuire, 2007; http://pyparsing.wikispaces.com) using the parse tree in Figure 2 (full code at https://github.com/JournalMap/JLIS_Geoparsers). PyParsing is recursive-descent parser that performs tokenization (i.e. identification of basic elements) and parsing as a single step rather than other commonly used parsers like Lex/Yacc that split tasks between two programs (Levine et al., 1995). This allows the grammar definition of the parser to include rules for both token assembly and syntax in an easy-to-read form rather than traditional approaches that start with a scanner to identify tokens using a definition file and then defines the syntax rules separately for a parser to use on the tokens. Additionally, PyParsing uses standard Python notation and does not rely on special character notation to define its grammars. Unlike the regular expression parser which scans text from left to right until it finds a character string that meets all of the syntax requirements, the PyParsing parser assigns identities to each token in a block of text based on the basic grammar rules. Only sets of contiguous tokens that can be built up into a coordinate according to the defined grammar are considered to be coordinates.

The coordinate parsers were developed iteratively using the grammar rules in Table 1. With each iteration of the parser, coordinates from the training set that were not parsed (omission errors) and coordinates that were parsed but did not match the JournalMap assigned coordinates were examined and used to modify the parser accordingly. This process continued until: (1) only incomplete or unusual coordinates were not parsed, and (2) all parsed coordinates matched their JournalMap-assigned locations. Examples of unusual coordinates that could not be correctly parsed included inconsistent formatting (e.g. use of minutes sign for seconds, inclusion of seconds for only one part of the coordinate pair) or non-standard notation (e.g. backslashes to separate minutes from seconds). The regular expression and lexical coordinate parsers correctly identified and extracted 98.7% and 98.5% of coordinates in the training set, respectively (Table 2). The regular expression and lexical coordinate parser code developed for this study is available on GitHub (https://github.com/JournalMap/Coordinateparsers).

Coordinate parser testing

When parsing geographic coordinates from article text, three types of errors can be made. Omission errors occur when the coordinate parser fails to find and extract a coordinate from an article. Commission errors occur when the coordinate parser mistakenly matches a set of numbers that are not really a coordinate. As examples, lists of morphological measurements or temperature ranges (especially those that do not denote temperature units) can be

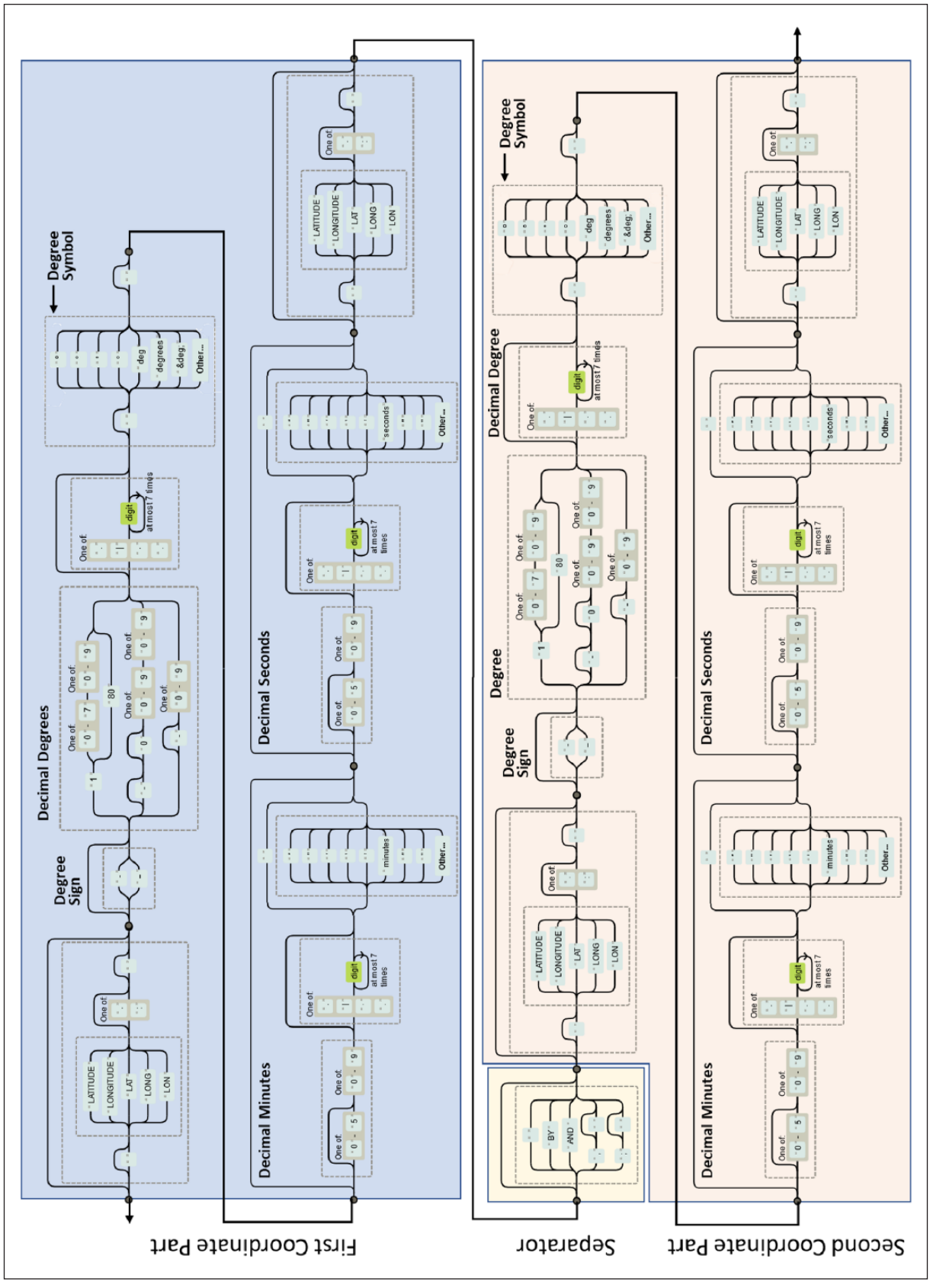

Figure 3. Syntax diagram showing how the regular expression parses geographic coordinates from text.
Note: Optional parts of a coordinate are indicated by a solid black line around an element. Dashed-line boxes indicate named groups in the regular expression. Names called out in the diagram denote elements used to calculate the coordinate. The regular expression was written to be able to parse coordinates in either latitude/longitude or longitude/latitude formats.

Table 2. Summary of coordinate parser performance with development coordinate set.

	Coordinate parser	
	Regular expression	Lexical
Locations parsed correctly	3079 (97.5%)	3051 (97.0%)
Locations parsed incorrectly	39 (1.2%)	48 (1.5%)
Poor or inconsistent formatting	24 (0.8%)	33 (1.0%)
Use of non-standard notation	15 (0.5%)	15 (0.5%)

confused for geographic coordinates. The third type of error, transformation error, occurs not in detecting and extracting coordinates from the article text, but in converting the coordinate pair to a standard decimal degree form.

For evaluating the performance of the coordinate parsers here, desired overall accuracy was at least 80% with omission error rates less than 15% and commission error rate less than 5%. While somewhat arbitrary, these evaluation goals are based on the following principles: (1) for coordinate parsing to result in a significant time savings in building databases of georeferenced literature, a high level of accuracy is required, and (2) not all error types have equivalent consequences. The principal cost of omission error in coordinate parsing is inefficiency – articles with coordinates that were missed would require use of another georeferencing technique (e.g. manual geotagging) or be omitted from searches. Conversely, commission errors can be very difficult to detect, and result in incorrect search results.

The approach described above for developing the coordinate parsers should result in robust parsers that can find, extract, and standardize many different variations of geographic coordinates. However, to evaluate the overall utility of the coordinate parsers, they must be tested on an independent set of articles. Also, because the training set of coordinates consisted only of coordinate pairs and occasionally a small amount of additional text, it was not possible to test the parsers for commission errors during the development phase and a separate test set was necessary.

To test the coordinate parsers, an independent set of 517 articles from seven different journals was assembled (Table 3). After manually checking each article, 141 articles (27.27%) were excluded because a study area was not appropriate (e.g. review articles, theoretical modeling research). Each of the remaining 376 articles in the test set was manually scanned for geographic coordinates and if found the coordinate pairs were copied verbatim from the article and stored for comparison to coordinate parser results. None of the test set of articles were used in developing the coordinate parsers or had previously been entered into the JournalMap database. Thus, their performance relative to the coordinate parsers was unknown.

Readers commonly interact with scientific articles in an online presentation (i.e. webpage) or as a Portable Document Format (PDF) file. However, most large publishers use an extensible markup language (XML) format for articles that facilitates transfer, archiving, and easy presentation in different formats. The National Library of Medicine (NLM) 3.0 (https://dtd.nlm.nih.gov/publishing) or NLM Journal Article Tag Suite (JATS) 1.0/1.1 (https://jats.nlm.nih.gov) formats are publishing-industry standard formats for storing published articles in a structured XML file. With these XML formats it easy to isolate specific article metadata (e.g. title, document object identifier [DOI]) or sections of the article text. For each article in the test set, full-text versions of the article in either the NLM or JATS format were obtained. Access to all the test articles was obtained either through agreements with the publishers of the journals or via open access licensing (PLoS ONE).

The regular expression and lexical coordinate parsers were run once over each article in the test set (i.e. no iteration to improve the coordinate parsers was done with the test set of articles), and the parsed coordinates were compared against the manually-obtained coordinate information for each article. Omission, commission, and transformation error rates were calculated for the test set. In cases where the parser made errors, the individual articles were examined to determine why the errors occurred, and each of the three different error types were summarized for each journal and for all journals combined.

Results

Proportion of articles in which a location was appropriate and incidence of geographic coordinates in articles varied by journal and publisher (Table 3). The natural history (*Journal of Natural History, Studies in Neotropical Fauna and the Environment*) and ecology journals (*South African Journal of Plant and Soil, The Auk, The Condor*) had the highest proportions of articles describing specific study areas and the highest rates of articles reporting geographic coordinates. *PLoS ONE* had the lowest proportion of articles that described a study area which is a reflection of the diverse nature of that journal. Articles from the Taylor & Francis journals had higher rates of reporting geographic coordinates (69.3% of location-appropriate articles from this publisher), due in part to efforts in cooperation with the JournalMap project to improve location reporting in those journals. For the other journals considered, geographic coordinates were reported in about half of studies (47.4% of location-appropriate articles). Incidence of

Table 3. Description of the set of articles used to test the performance of the coordinate parsers.

Journal	Publisher	Year range	Total articles evaluated	Articles where location is appropriate	Articles reporting point geographic coordinates[a]	Articles reporting extent or bounding box[a]
PLoS ONE	PLOS	2007–2016	203	121 (59.6%)	56 (46.3%)	23 (19.0%)
Journal of Natural History	Taylor & Francis (T&F)	2015–2016	58	54 (93.1%)	32 (59.3%)	4 (7.4%)
South African Journal of Plant and Soil	Taylor & Francis	2015–2016	9	8 (88.9%)	7 (87.5%)	0 (0%)
Studies in Neotropical Fauna and the Environment	Taylor & Francis	2015–2016	37	36 (97.3%)	33 (91.7%)	1 (2.8%)
The Condor	American Ornithologists' Union (AOU)	2014	53	40 (75.5%)	17 (42.5%)	1 (2.5%)
The Auk	American Ornithologists' Union	2014	60	50 (83.3%)	26 (52%)	3 (6.0%)
AoB PLANTS	Oxford University Press (OUP)	2009–2013	97	67 (69.1%)	33 (49.3%)	2 (3.0%)

[a]Calculated as the number and percentage of articles for which a location was appropriate.

reporting bounding boxes instead of simple point coordinates was generally low for all journals except *PLoS ONE*.

Both coordinate parsers performed well at recognizing and extracting coordinates from articles across publishers (Table 3, Figure 4). Total percent correct for the regular expression parser ranged from 85.1% for *Journal of Natural History* to 100% for *The Condor*. The lexical parser performed slightly better with total percent correct ranging from 88.9% for *Journal of Natural History* to 100% for four journals. In all cases, coordinates correctly identified and extracted with the regular expression parser were also correctly extracted by the lexical parser. In all articles considered for this study, there was only a single commission error which occurred with the lexical parser for one *PLoS ONE* article. In this study, the authors give a description of plot transect configuration that appears similar to a geographic coordinate: "We established survey locations at the patch center or grassland site 'center' and 50 m from those points at 0°, 120°, and 240°" (Gould et al., 2013).

Average number of locations per article was similar between the regular expression and lexical coordinate parsers with the exception of the *Journal of Natural History* (Figure 4). This journal, which publishes species descriptions and often reports coordinates for type specimens, had a much higher number of locations per article than any other journal, and the PyParsing algorithm was more efficient (in terms of number of locations found) at detecting and extracting these locations. In general, though, the lexical coordinate parser detected more coordinates and yielded a greater geographic distribution of study locations that the regular expression coordinate parser (Figure 5).

However, both coordinate parsers were prone to recognizing bounding boxes as point coordinates depending on how they were formatted. Bounding boxes that are presented as either point coordinates that define the corners or the box (e.g. 18°37′16N–92°42′28W to 18°30′20N–91°28′03W, Villéger et al., 2012) or as latitude/longitude ranges without a clear distinction between the parts (e.g. 127°42′55″–128°16′48″E, 41°41′49″–42°25′18″N, Zhang et al., 2014) are interpreted by the coordinate parsers as point coordinates. This occurred 21 times (44.1% of all articles reporting bounding boxes).

Omission error rates for the lexical coordinate parser were equal to or lower than those of the regular expression coordinate parser for all journals (Table 4). Over a third of omission errors (38.5%) were the result of coordinate errors (e.g. latitudes greater than 90º) or poorly formatted coordinates (e.g. missing degree signs or use of unusual characters) in the journal articles (Table 5). Locations described by bounding boxes were the second largest source of omission errors (35.8%) (Table 5). Point coordinates reported in tables or article supplements accounted for 15.4% of omission errors. In these cases, coordinates were missed if the coordinate pairs were split into different rows or columns or if notation such as degree signs were omitted. With four articles (10.4% of all omission errors), there was no identifiable cause for the omission error.

Discussion

The coordinate parsers developed and evaluated here demonstrate that it is feasible to accurately detect and extract geographic coordinates from published studies and use this information to automate georeferencing of articles based on where the research was conducted. Both coordinate parsers detected coordinates that appeared in articles across publishers with minimal errors, with the lexical coordinate parser performing marginally better.

Detecting coordinates in published articles is, to some degree, an exercise in optimization. If the coordinate parser is defined too rigidly, then coordinates that appear in a study may be missed (i.e. omission errors). Conversely,

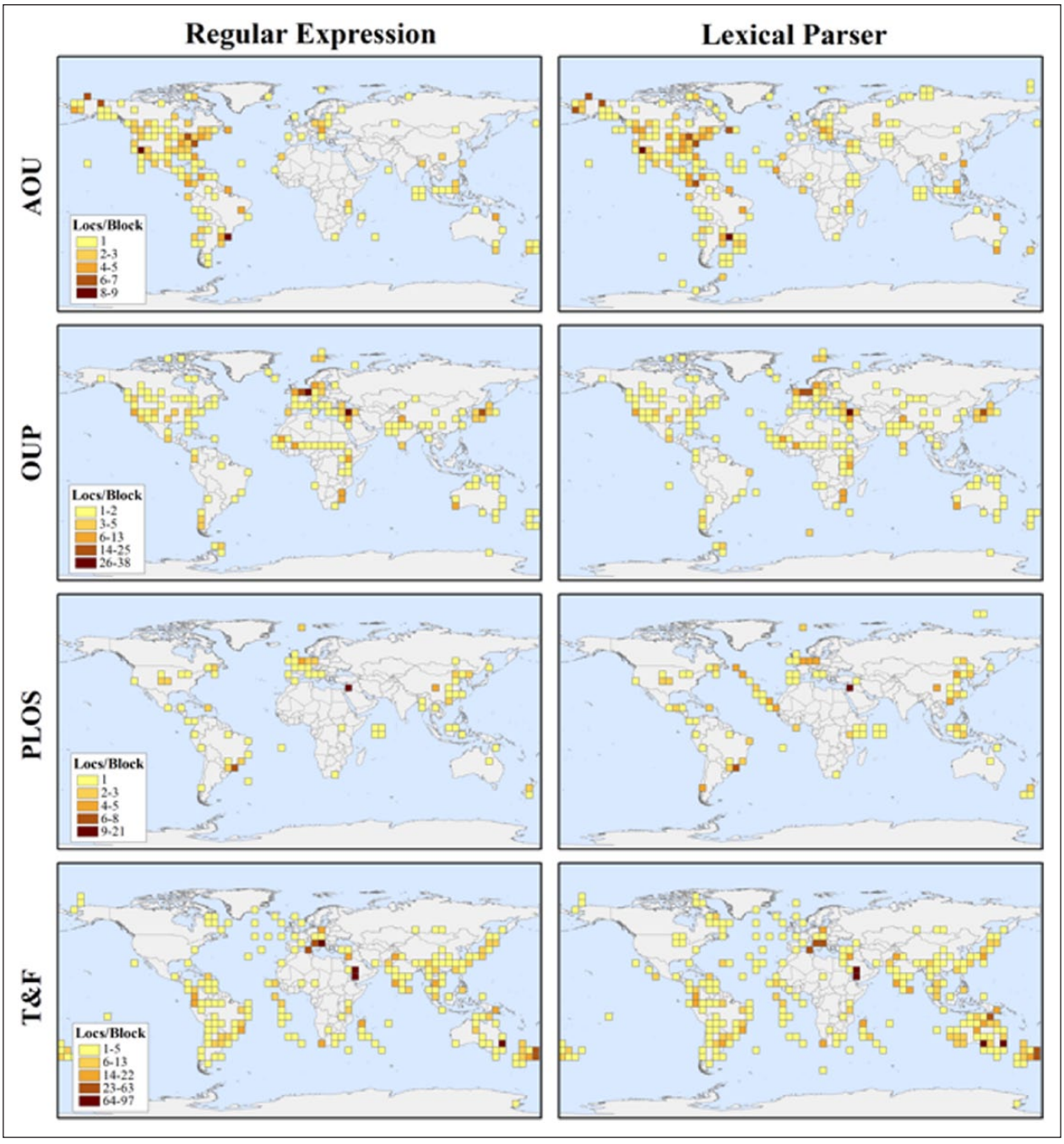

Figure 4. Maps of the number of study locations extracted from 517 test articles by publisher and coordinate parser.
Note: While the total number of articles from which coordinates were extracted was similar among the coordinate parser types, the lexical coordinate parser tended to identify more coordinates per article than did the regular expression version. See Table 3 for details on how many articles from each journal had coordinates that could be recognized and extracted. Publisher abbreviations are given in Table 3.

if the parser is defined too loosely (e.g. relaxing the requirement of degree signs), then sequences of numbers may be misinterpreted as geographic coordinates (i.e. commission errors). In the context of automating the identification and extraction of location information from scientific literature, commission errors may be the more costly error type because there is little opportunity to detect them. Omission errors, on the other hand, may be dealt with by re-mining the article sets as parsers improve.

Despite the coordinate parsers performing very well at recognizing, extracting, and standardizing geographic coordinates in published articles, the lack of conventions for how study areas are reported and formatted in a scientific article resulted in many location-appropriate articles not being georeferenced. This was either because they lacked any coordinates or because of formatting issues with or errors in coordinates that were reported. The extreme variability in how locations are reported in articles

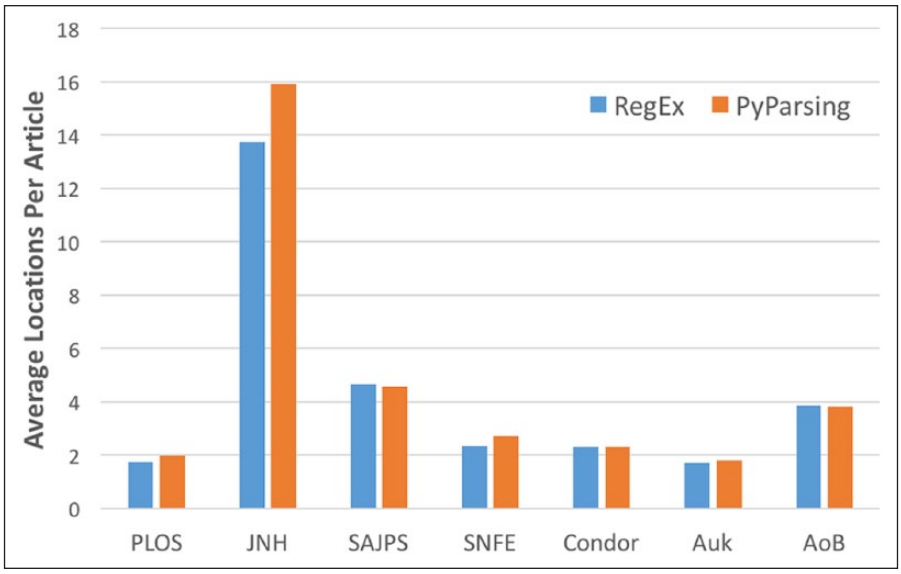

Figure 5. Average number of locations per article that were recognized and extracted by the RegEx and PyParsing coordinate parsers.
Note: PLOS=PLoS ONE, JNH=Journal of Natural History, SAJPS=South African Journal of Plant and Soil, SNFE=Studies in Neotropical Fauna and the Environment, AoB=AoB PLANTS.

Table 4. Performance of the regular expression and lexical coordinate parsers for the test set of articles from seven journals.

Journal	Regular expression						Lexical					
	Correct Coord.	Correct None	Total % Correct	Om. Error	% Om. Error	Com. Error	Correct Coord.	Correct None	Total % Correct	Om. Error	% Om. Error	Com. Error
PLoS ONE (*n*=121)	66	37	85.1%	18	14.9%	0	73	36	90.1%	11	9.1%	1
The Auk (*n*=50)	24	21	90%	5	10.0%	0	26	24	100%	0	0.0%	0
The Condor (*n*=40)	18	22	100%	0	0.0%	0	18	22	100%	0	0.0%	0
AoB PLANTS (*n*=67)	29	33	92.5%	5	7.5%	0	29	33	92.5%	5	7.5%	0
South African Journal of Plant and Soil (*n*=8)	6	1	87.5%	1	12.5%	0	7	1	100%	0	0.0%	0
Studies in Neotropical Fauna and the Environment (*n*=36)	33	2	97.2%	1	2.8%	0	34	2	100%	0	0.0%	0
Journal of Natural History (*n*=54)	29	17	85.2%	8	14.8%	0	31	17	88.9%	6	11.1%	0

Note: Total percent correct was calculated as the sum of the number of articles correctly identified as having a coordinate (Correct Coord.) and articles correctly identified as not having a coordinate (Correct None) divided by the total number from that journal. Omission error (Om. Error, coordinates in an article that were missed) rate was calculated as the number of articles in which an omission error occurred, and % omission error was the number of omissions divided by the total number of articles. Commission errors (Com. Error) occurred from incorrectly identifying a character sequence as being a coordinate.

Table 5. Explanation of omission errors in geoparsing the test set of articles from six journals.

	Bounding box	Coordinates in table or supplement	Coordinate formatting or error	Unknown
PLoS ONE	11	1	6	0
The Auk	1	2	2	0
AoB PLANTS	2	0	3	0
South African Journal of Plant and Soil	0	1		0
Studies in Neotropical Fauna and the Environemnt	0	1	1	0
Journal of Natural History	0	1	3	4
TOTAL[a]	14 (35.8%)	6 (15.4%)	15 (38.5%)	4 (10.3%)

[a]Percentage omission error calculated as number of omissions by explanation divided by the total number of omissions across all journals (*n*=39).

and the resulting challenges this poses for accurately and reliable georeferencing these studies supports an argument for standard approaches to including location information in manuscripts (Page, 2010).

Text-based descriptions of study areas are almost universal for articles describing field-based primary research, but reporting geographic coordinates that describe a point (e.g. centroid) in the study area, the corners of a bounding box, or a latitude or longitude zone occurs in only about half of ecology articles (Karl et al., 2013). This means that automated georeferencing of articles using parsers as described here can capture only approximately half of the available articles. The remainder must be either manually georeferenced or extracted using place-name recognition algorithms. For many of these articles, however, coordinates could easily be added to describe the study area if it were encouraged or required by the journals. Prevalence of geographic coordinates in articles for the three Taylor & Francis journals considered here increased markedly (from an average of 45% to 56% of articles including coordinates, JournalMap, unpublished data) within two months after the publisher added a geolocation reporting section to their instructions for authors (https://journalmap.org, unpublished data). Given the challenges discussed above of georeferencing articles from place-names, even if (or when) robust place-name recognition algorithms can be implemented, coordinates in an article could serve to validate location information.

The myriad ways in which geographic coordinates are stylistically formatted in journals complicates geoparsing. For example, to achieve the reported level of accuracy with the coordinate parsers presented here required nine different symbols or words for a degree sign. Even within the same journal, formatting of coordinates may differ dramatically. Another challenge to geoparsing is formatting that breaks the semantic structure used to identify a sequence of characters as a coordinate. Particularly difficult are coordinates reported in tables where the parts of the coordinate pair are split into separate columns or rows. Coordinates reported in supplementary material may also be difficult to parse because supplements may not be included in standard article XML or formatted consistently. In the absence of standards for reporting location information, decisions on formatting coordinates are typically left to the discretion of the authors.

While almost all articles considered for this study reported coordinates using some combination of degrees, minutes, and seconds, locations are occasionally reported using other coordinate systems (e.g. UTM, Military Grid Reference System). In most cases choice of an uncommon coordinate system was purely author preference. However, the use of uncommon coordinate systems is sometimes justified and likely varies by journal or science field. For example, articles in polar research journals may be more likely to report coordinates using polar stereographic coordinates due to shortcomings of degree-based geographic coordinates at the poles. It is certainly possible to write coordinate parsers for any well-defined coordinate system, but challenges may arise if the syntax of a coordinate system is not well distinguished from other commonly-reported data types. For example, Page (2010) found that GenBank identifiers were easily confused with UTM grid references (e.g. DQ402119). Additionally, formatting of bounding boxes caused confusion in the point geoparsers. This could be addressed through a separate coordinate parser to detect bounding boxes that would be applied before parsing point coordinates from an article.

Most of the omission errors encountered in geoparsing the articles for this study were the result of either coordinate errors or poor coordinate formatting. The JournalMap project has found approximately 2 to 5% of studies (depending on the journal) that report coordinates for a location have obvious errors such as invalid values (e.g. latitudes greater than 90º), improper formatting (e.g. the use of both a negative sign and "W" to specify hemisphere direction for longitude), incomplete coordinates (e.g. Universal Transverse Mercator coordinates that do not give the zone) (unpublished data). The actual rate of errors in reported coordinates is likely higher than this, though, as these parsing approaches can detect only obvious errors. Coordinate errors can originate in the submitted manuscript, or can be introduced in the formatting and typesetting process. For example, Bestelmeyer et al. (2006) reported their study area location correctly using UTM coordinates, but the coordinate values were erroneously reformatted (without applying a geographic transformation) into degrees, minutes, and seconds by the publisher, resulting in the coordinate describing a location in the middle of the Mediterranean Sea that should have been in the southwestern deserts of the United States. This error is known only by knowledge of the study area and personal communication with the authors. The important point here, however, is that this error was not caught during review of the article proofs and was subsequently published. All these issues highlight the fact that location information is not being verified during the manuscript review and production process. The adoption of a standard for reporting study locations in articles could help in these cases through providing a consistent and predictable means for describing study locations. Reporting standards could also support tools for validating locations during the peer-review process such as maps for the authors to verify correct study area locations.

Attributes of a location reporting standard for scientific studies

One potential solution to alleviate some of the challenges of extracting geographic coordinates from articles is to encode study area locations into article metadata using existing mechanisms. Locations formally encoded into article metadata using a standard set of elements or tags could then be read directly rather than needing to be parsed from unstructured text. For example, the Dublin Core document metadata standard already includes a Coverage element to be used for

describing the "spatial and temporal extent of [an] object or resource and is the key element for supporting spatial or temporal range searching on document-like objects…" (http://dublincore.org/). Within the Dublin Core protocol, geographic locations can be defined as points, lines, bounding boxes, and polygons as well as via place names (with reference to a published place-name Gazeteer). As part of the article metadata, automated geolocation of articles via place names is possible because the place name that describes the study is contained within a metadata element that is easily identifiable, and the reference to a Gazeteer makes the place name unambiguous. As an example of this, the publisher Pensoft (http://pensoft.net/journals) encodes geographic locations in their articles using the Darwin Core metadata protocol (Wieczorek et al., 2012; http://rs.tdwg.org/dwc/), a derivation of the Dublin Core for biodiversity literature.

Regardless of whether study area locations are reported in the article text or in the article's metadata, standards for when and how locations will be reported would greatly improve the ability to geotag and use location information from articles. Study area descriptions should be reported according to the proposed format for all primary research that reports on data collected or observations made at one or more locations on the ground or that draws inferences to a definable area or location. Additionally, it may be appropriate for some secondary research studies (e.g. meta-analyses) to report study locations according to the proposed standard when they synthesize primary research studies with the intent of drawing inferences to a definable area or location. Study locations should be described in enough detail that they can be accurately mapped at an appropriate scale for the study A concise text-based description should be included even when coordinates are reported to allow for verification of coordinate values. Additionally, the following principles should be employed in development of location reporting standards for describing study areas in scientific research.

- *Provide complete, accurate, and precise location information* – Locations described using the standard should be able to be accurately (i.e. without error or ambiguity) and precisely located on a map. Study area descriptions should contain enough information to permit the identification, use, and conversion of location information without the reliance on external resources.
- *Provide location information at a scale that is appropriate to the study* – The location reporting standard should allow for the study area to be described at a scale that is appropriate to the study. The standard should document the inference or focus area of a study and not just the individual observations used in the study. For example, a point coordinate may adequately describe the location of a study from a specific research site, but a study of a large landscape or region may be better described by a bounding box.
- *Be flexible* – A location standard should allow for use of coordinate or place reference systems that are appropriate for describing the study area and compatible with the needs and conventions of different fields of study.
- *Use widely-used place-name and coordinate formats* – A location reporting standard should be built on formats that are broadly implemented so that existing tools and technologies can be used to map, extract, convert, and analyze locations. If an alternative to the geographic coordinate system is justified (e.g. polar regions), all necessary information is provided (e.g. zone numbers).
- *Be extensible* – A location reporting standard should include provisions for modification and additions as appropriate.
- *Not be easily confused with other data types or presentations* – Formatting of geographic coordinates or locations should be distinct from other commonly reported standard data types or references (e.g. specimen or material identifiers, accession numbers, catalog reference numbers) to minimize commission errors.
- *Produce machine readable location information* – Location information included according to the standard should be able to be easily recognized and extracted from text and converted to other formats by computer algorithms.

The two coordinate parsers tested each had their strengths and weaknesses. The lexical parser detected coordinates in more articles and found more coordinates per article, but was computationally intensive and slower than the regular expression version. The regular expression parser was much quicker but did not perform as well and would be challenging to maintain owing to its complex and hard to read structure. A hybridized implementation where articles are scanned first by a simplified regular expression coordinate parser and if no coordinates were detected then scanned again by a more robust lexical coordinate parser may be a good solution for offsetting the weakness of each parser.

Ideally, a robust system for article geotagging would include coordinate parsers as described here as well as named entity or toponym recognition algorithms for identifying and geocoding place names. For this to be effective in georeferencing the study locations scientific articles, however, effective strategies for filtering out irrelevant place names must be developed. Ranking algorithms have been explored for determining the geographic scope of documents such as web pages or news stories that may contain ambiguous or irrelevant location information (e.g. Monteiro et al., 2016; Silva et al., 2006). Such ranking approaches rely on rules (e.g. place names mentioned more often are more likely to be relevant) and contextual cues to filter extraneous locations and determine the geographic scope of a document. This kind of approach may work for georeferencing scientific articles and

should be explored, but may require rulesets specific to scientific literature.

Conclusion

Despite the differences between the coordinate parsers in terms of how many coordinates they detected, the main conclusion from this study is that standard parsing tools can be used successfully to identify and extract a wide variety of geographic coordinates from scientific literature. While it is certainly possible to develop better parsers for geographic coordinates (e.g. to handle alternate coordinate systems, bounding boxes, inconsistent formatting) this effort may yield only marginal gains in article georeferencing. Instead, approaches that marry geographic coordinate detection with place name recognition may be more fruitful for creating automated systems for article georeferencing. Ultimately, though, the success of geoparsing locations from published articles will hinge upon whether better standards for location reporting can be implemented in scholarly publishing. The adoption of standards for location reporting and encoding of location information in machine-readable forms within article metadata will amplify the ability for existing scientific literature to be found and used to support novel research.

The notions of georeferenced bibliography or location-based searching for literature are not new, but have previously been hampered by the effort needed to manually georeference articles, a process that is onerous, time-consuming, and error prone. With the growing support for deep text and data mining of article content by most large publishers, there is a tremendous new opportunity to achieve a scale of literature georeferencing that could make map-based searching truly useful across a wide range of disciplines. Reliable, automated techniques like coordinate parsers for assigning study locations to scientific articles could contribute to this which, in turn, could help map-based literature searching become standard practice for scientific research. The increased ability to discover research from specific (or similar) places could help reduce redundancy in conducting new studies, aid in knowledge transfer to understudied regions, promote novel syntheses and meta-analyses of existing research, uncover new patterns to systems and events across the world, and encourage study of bias in scientific research.

Acknowledgements

I would like to acknowledge the assistance of R Baca, and JA Karl in assembling the article test sets used in this research. The American Ornithologists' Union, Oxford University Press, and Taylor & Francis graciously allowed content from several of their journals to be used in this study. J Gillan, J Maynard, and R Salzman provided valuable comments and edits on an earlier draft of this study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by appropriated funding to the USDA-ARS Rangeland Management Research Unit, Jornada Experimental Range.

ORCID iD

Jason W Karl https://orcid.org/0000-0002-3326-3806

References

Aho AV, Sethi R and Ullman JD (1986) *Compilers, Principles, Techniques, and Tools*. Reading, MA: Addison-Wesley.

Amitay E, Har'El N, Sivan R, et al. (2004) Web-a-where: Geotagging web content. In: *27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '04)*, Sheffield, UK, 25–29 July 2004, pp. 273–280. New York: ACM. Available at: http://portal.acm.org/citation.cfm?doid=1008992.1009040 (accessed 25 July 2016).

Bestelmeyer BT, Ward JP, Herrick JE, et al. (2006) Fragmentation effects on soil aggregate stability in a patchy arid grassland. *Rangeland Ecology & Management* 59(4): 406–415.

Borges KAV, Davis CA, Laender AHF, et al. (2011) Ontology-driven discovery of geospatial evidence in web pages. *GeoInformatica* 15(4): 609–631.

Fisher R, Radford BT, Knowlton N, et al. (2011) Global mismatch between research effort and conservation needs of tropical coral reefs. *Conservation Letters* 4: 64–72.

Ford B (2004) Parsing expression grammars: A recognition-based syntactic foundation. In: *31st ACM SIGPLAN-SIGACT symposium on principles of programming languages (POPL '04)*, Venice, Italy, 14–16 January 2004, pp. 111–122. New York: ACM. Available at: http://portal.acm.org/citation.cfm?doid=964001.964011 (accessed 31 May 2016).

Friedl JEF (2006) *Mastering Regular Expressions*. 3rd edn. Sebastapol, CA: O'Reilly.

Gould RK, Pejchar l, Bothwell SG, et al. (2013) Forest restoration and parasitoid wasp communities in Montane Hawai'i. *PLoS ONE* 8(3): e59356.

Grune D and Jacobs CJ (1990) *Parsing Techniques: A Practical Guide*. Ellis Horwood Series in Computers and their Applications. New York: Horwood.

Hendrickx D, Dujardin J-C, Pickering J, et al. (2010) The leishmaniasis e-compendium: A geo-referenced bibliographic tool. *Trends in Parasitology* 26(11): 515–516.

Hughes TP, Baird AH, Dinsdale EA, et al. (2002) Detecting regional variation using meta-analysis and large-scale sampling: Latitudinal patterns in recruitment. *Ecology* 83: 436–451.

Jetz W, McPherson JM and Guralnick RP (2012) Integrating biodiversity distribution knowledge: Toward a global map of life. *Trends in Ecology & Evolution* 27(3): 151–159.

Johnson WL, Porter JH, Ackley SI, et al. (1968) Automatic generation of efficient lexical processors using finite state techniques. *Communications of the ACM* 11(12): 805–813.

Jones CB and Purves RS (2008) Geographical information retrieval. *International Journal of Geographical Information Science* 22: 219–228.

Jones CB, Purves R, Ruas A, et al. (2002) Spatial information retrieval and geographical ontologies: An overview of the SPIRIT project. In: *25th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '02)*, Tampere, Finland, 11–15 August 2002, pp. 387–388. New York: ACM. Available at: http://portal.acm.org/citation.cfm?doid=564376.564457 (accessed 25 July 2016).

Karl JW, Gillan JK and Herrick JE (2013) Geographic searching for ecological studies: A new frontier. *Trends in Ecology & Evolution* 28(7): 383–384.

Karl JW, Herrick JE, Unnasch RS, et al. (2013) Geo-semantic searching: Discovering ecologically-relevant knowledge from published studies. *BioScience* 63(8): 674–682.

Leidner JL (2007) *Toponym resolution in text: Annotation, evaluation and applications of spatial grounding of place names*. Boca Raton, FL: Dissertation.com.

Levine JR, Mason T and Brown D (1995) *Lex and Yacc: UNIX Programming Tools. A Nutshell Handbook*. 2nd edn. [repr. with minor corr]. Sebastopol, CA: O'Reilly.

Lieberman MD and Samet H (2011) Multifaceted toponym recognition for streaming news. In: *34th international ACM SIGIR conference on research and development in information retrieval (SIGIR '11)*, Beijing, China, 24–28 July 2011, pp. 843–852. New York: ACM. Available at: http://portal.acm.org/citation.cfm?doid=2009916.2010029 (accessed 3 June 2017).

Lieberman MD, Samet H and Sankaranarayanan J (2010) Geotagging with local lexicons to build indexes for textually-specified spatial data. In: *26th international conference on data engineering (ICDE)*, Long Beach, California, USA, pp. 201–212. IEEE. Available at: http://ieeexplore.ieee.org/document/5447903/ (accessed 3 June 2017).

McGuire P (2007) *Getting Started with PyParsing*. Sebastopol, CA: O'Reilly. Available at: http://public.eblib.com/choice/publicfullrecord.aspx?p=3027028 (accessed 1 June 2016).

Maggio A, Kuffer J and Lazzari M (2017) Advances and trends in bibliographic research: Examples of new technological applications for the cataloguing of the georeferenced library heritage. *Journal of Librarianship and Information Science* 49(3): 299–312.

Magliocca NR, Rudel TK, Verburg PH, et al. (2014) Synthesis in land change science: Methodological patterns, challenges, and guidelines. *Regional Environmental Change*. Available at: http://link.springer.com/10.1007/s10113–014–0626–8 (accessed 6 November 2014).

Martin LJ, Blossey B and Ellis E (2012) Mapping where ecologists work: Biases in the global distribution of terrestrial ecological observations. *Frontiers in Ecology and the Environment* 10: 195–201.

Monteiro BR, Davis CA and Fonseca F (2016) A survey on the geographic scope of textual documents. *Computers & Geosciences* 96: 23–34.

Page RD (2011) Extracting scientific articles from a large digital archive: BioStor and the Biodiversity Heritage Library. *BMC Bioinformatics* 12(1): 187.

Page RDM (2010) Enhanced display of scientific articles using extended metadata. *Web Semantics: Science, Services and Agents on the World Wide Web* 8(2/3): 190–195.

Schmill MD, Gordon LM, Magliocca NR, et al. (2014) GLOBE: Analytics for assessing global representativeness. In: *Fifth international conference on computing for geospatial research and application*, Washington DC, USA, 4–6 August 2014, pp. 25–32. IEEE. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6910113 (accessed 6 November 2014).

Schmitt J and Butler B (2012) Creating a geo-referenced bibliography with Google Earth and GeoCommons: The Coos Bay Bibliography. *Issues in Science and Technology Librarianship* 71(Fall 2012).

Shapiro JT and Báldi A (2012) Lost locations and the (ir)repeatability of ecological studies. *Frontiers in Ecology and the Environment* 10: 235–236.

Silva MJ, Martins B, Chaves M, et al. (2006) Adding geographic scopes to web resources. *Computers, Environment and Urban Systems* 30(4): 378–399.

Smith DA and Crane G (2001) Disambiguating geographic names in a historical digital library. In: Constantopoulos P and Sølvberg IT (eds) *Research and Advanced Technology for Digital Libraries*. Berlin, Heidelberg: Springer, pp. 127–136. Available at: http://link.springer.com/10.1007/3–540–44796–2_12 (accessed 25 July 2016).

Thompson K (1968) Programming techniques: Regular expression search algorithm. *Communications of the ACM* 11(6): 419–422.

van Vliet N, Mertz O, Heinimann A, et al. (2012) Trends, drivers and impacts of changes in swidden cultivation in tropical forest-agriculture frontiers: A global assessment. *Global Environmental Change* 22(2): 418–429.

Villéger S, Miranda JR, Hernandez DF, et al. (2012) Low functional β-diversity despite high taxonomic β-diversity among tropical estuarine fish communities. *PLoS ONE* 7(7): e40679.

Volz R, Kleb J and Mueller W (2007) Towards ontology-based disambiguation of geographical identifiers. In: *16th international World Wide Web conference (WWW 2007)*, Banff, Alberta, Canada, 8–12 May 2007. Available at: https://www.researchgate.net/publication/220718075_Towards_Ontology-based_Disambiguation_of_Geographical_Identifiers (accessed 31 January 2018).

Wieczorek J, Bloom D, Guralnick R, et al. (2012) Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE* 7(1): e29715.

Zhang J, Liu F and Cui G (2014) The efficacy of landscape-level conservation in Changbai Mountain Biosphere Reserve, China. *PLoS ONE* 9(4): e95081.

Zimmerman A (2007) Not by metadata alone: The use of diverse forms of knowledge to locate data for reuse. *International Journal on Digital Libraries* 7: 5–16.

Author biography

Jason W Karl is an Associate Professor and Harold F and Ruth M Heady Endowed Chair of Rangeland Ecology in the Department of Forest, Rangeland, and Fire Sciences at the University of Idaho. Dr Karl created the JournalMap geographic literature search project while a Research Ecologist with the USDA Agricultural Research Service.